Traffic congestion and congestion pricing

Verhoef, E.T.; Lindsey, C.R.

2000

TI 2000-101/3
Tinbergen Institute Discussion Paper

Traffic Congestion and Congestion Pricing

C. Robin Lindsey
Erik T. Verhoef

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

TRAFFIC CONGESTION AND CONGESTION PRICING*

C. Robin Lindsey[1] and Erik T. Verhoef[2**]

[1]Department of Economics
University of Alberta
Edmonton
Alberta
Canada
Phone: +1-780-4927642
Fax: +1-780-4923300
E-mail: crli@econ.ualberta.ca

[2]Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
Phone: +31-20-4446094
Fax: +31-20-4446004
E-mail: everhoef@econ.vu.nl

This version: 14/11/00

Key words: congestion, road pricing, networks

JEL codes: R41, R48, D62

Abstract

For several decades growth of traffic volumes has outstripped investments in road infrastructure. The result has been a relentless increase in traffic congestion. This paper reviews the economic principles behind congestion pricing in static and dynamic settings, which derive from the benefits of charging travellers for the externalities they create. Special attention is paid to various complications that make simple textbook congestion pricing models of limited relevance, and dictate that congestion pricing schemes be studied from the perspective of the theory of the second best. These complications include pricing in networks, heterogeneity of users, stochastic congestion, interactions of the transport sector with the rest of the economy, and tolling on private roads. Also the implications of congestion pricing for optimal road capacity are considered, and finally some explanations for the longstanding social and political resistance to road pricing are offered.

*The authors would like to thank Richard Arnott, André de Palma, Claude Penchina, Stef Proost and especially Ken Small for corrections and insightful comments on an earlier version of this paper. Any remaining errors, however, are the authors' responsibility alone.

**Erik Verhoef is affiliated as a research fellow to the Tinbergen Institute. Erik Verhoef's research has been supported by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

1. Introduction

For several decades growth of traffic volumes has outstripped investments in road infrastructure. The result has been a relentless increase in traffic congestion. Congestion imposes various costs on travellers: reduced speeds and increased travel times, a decrease in travel time reliability, greater fuel consumption and vehicle wear, inconvenience from rescheduling trips or using alternative travel modes, and (in the longer run) the costs of relocating residences and jobs. The costs of increased travel times and fuel consumption alone are estimated to amount to hundreds of dollars per capita per year in the US (Schrank and Lomax, 1999) and comparable values have been reported for Europe.

Traffic congestion is a consequence of the nature of supply and demand: capacity is time-consuming and costly to build and is fixed for long time periods, demand fluctuates over time, and transport services cannot be stored to smooth imbalances between capacity and demand. Various policies to curb traffic congestion have been adopted or proposed over the years. The traditional response is to expand capacity by building new roads or upgrading existing ones. A second method is to reduce demand by discouraging peak-period travel, limiting access to congested areas by using permit systems and parking restrictions, imposing bans on commercial vehicles during certain hours, and so on. A third approach is to improve the efficiency of the road system, so that the same demand can be accommodated at a lower cost. Re-timing of traffic lights, metering access to highway entrance ramps, high-occupancy vehicle lanes and Advanced Traveller Information Systems are examples of such measures.

This paper is concerned with congestion pricing as a tool for alleviating traffic congestion. The insight for congestion pricing comes from the observation that people tend to make socially efficient choices when they are faced with all the social benefits and costs of their actions. As just noted various demand management tools to accomplish this can be used. But congestion pricing is widely viewed by economists as the most efficient means because it employs the price mechanism, with all its advantages of clarity, universality, and efficiency. Pigou (1920) and Knight (1924) were the first to advocate it. But it was the late William Vickrey, who steadfastly promoted congestion pricing for some forty years, who was arguably the most influential in making the case on both theoretical and practical grounds. In one of his early advocacy pieces, Vickrey (1963) identified the potential for road pricing to influence travellers' choice of route and travel mode, and its implications for land use. He also discussed alternative methods of automated toll collection. Another of his early proposals was to set parking fees in real time as a function of the occupancy rate. An overview of Vickrey's contributions to pricing of urban private and public transport is found in Arnott *et al*. (1994, pp. 271-5).

As Vickrey's work makes clear, true congestion pricing entails setting tolls that match the severity of congestion, which requires that tolls vary according to time, location, type of vehicle and current circumstances (e.g. accidents or bad weather). Congestion pricing is

common in other sectors of the economy — from telephone rates and air fares to hotels and public utilities. But despite the efforts of Vickrey and other economists, congestion pricing is still rarely used on roads. Tolls are not charged on most roads, and fuel taxes do not vary with traffic volumes. And costs of registration, licensing and insurance do not even depend on distance travelled. Nevertheless, the number of applications and experiments in road pricing is slowly growing, spurred on by the combined impetus of worsening traffic conditions and advances in automatic vehicle identification technology. Descriptions of various road pricing schemes, including Singapore's pioneering toll system, Scandinavian toll-rings, and Californian pay-lanes are found in Gómez-Ibáñez and Small (1994) and Small and Gómez-Ibáñez (1998).

This paper is organized along similar lines to the review of congestion modelling in Lindsey and Verhoef (2000). Section 2 starts by outlining the basic economic principles of congestion pricing in a simple static ('time-independent') setting with one road. Section 3 adds a time element by considering travellers' time-of-use decisions and time-varying tolls. Various complications are addressed in Section 4, including pricing in networks, heterogeneity of users, stochastic congestion, interactions of the transport sector with the rest of the economy, and tolling on private roads. Section 5 considers the implications of congestion pricing for optimal road capacity. Explanations for the longstanding social and political resistance to road pricing are offered in Section 6, and conclusions are drawn in Section 7. Due to space constraints some topics related to congestion pricing are not covered in this review. There is no explicit treatment of freight transportation. Nothing is said about the implications of congestion pricing for urban structure or the location of new developments. And only passing mention in Section 4 is given to the potential effects of congestion pricing on traffic noise, pollution, and traffic accidents.

2. Congestion pricing in time-independent models

The basic principles of congestion pricing can be illustrated in the following simple setting. Consider one origin and one destination connected by a single road. Individuals make trips alone in identical vehicles. Traffic flows, speeds and densities are uniform along the road and independent of time. Equilibrium in this setting is described in Figure 1, which is due to Walters (1961). The horizontal axis depicts traffic flow or volume: the rate at which trips are initiated and completed. The vertical axis depicts the price or 'generalized cost' of a trip — which includes vehicle operating costs, the time costs of travel, and any toll. At low volumes vehicles can travel at the free-flow speed, and the trip cost curve, $C(q)$, is constant at the free-flow cost C^{ff}. At higher volumes congestion develops, speed falls, and $C(q)$ slopes upwards. (Figure 1 ignores the possibility of 'hypercongestion' that would cause $C(q)$ to bend backwards on itself; see Lindsey and Verhoef, 2000.)

If flow is interpreted to be the quantity of trips "demanded" per unit of time, then a demand curve $p(q)$ can be added to Figure 1 to obtain a supply-demand diagram. The demand

curve is assumed to slope downwards to reflect the fact that, as for most commodities, the number of trips people want to make decreases with the price. The unregulated 'no-toll' equilibrium occurs at the intersection of $C(q)$ and $p(q)$, resulting in an equilibrium flow of q^n and an equilibrium price of C^n. Since 'external benefits' of road use are not likely to be significant (benefits are normally either purely internal or pecuniary in nature), $p(q)$ specifies both the private and the marginal social benefit of travel. Total social benefits can thus be measured by the area under $p(q)$. Analogously, $C(q)$ measures the cost to the traveller of taking a trip. If external travel costs other than congestion, such as accidents and air pollution, are ignored, then $C(q)$ measures the average social cost of a trip. The total social cost of q trips is then $TC(q) = C(q) \cdot q$, and the marginal social cost of an additional trip is $MC(q) = \partial TC(q)/\partial q = C(q) + q \cdot \partial C(q)/\partial q$.

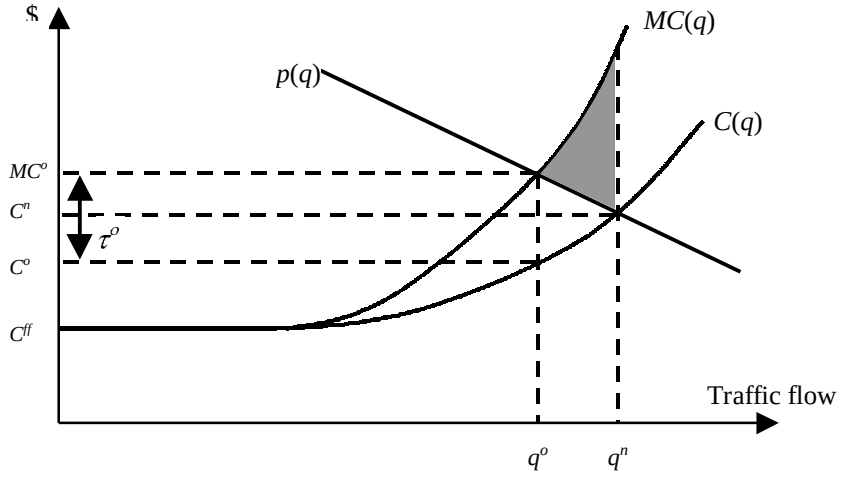

Figure 1. Optimal road pricing in a time-independent model

The social optimal is found in Figure 1 at the intersection of $MC(q)$ and $p(q)$, where the marginal willingness to pay for trips is MC^o and the number of trips, q^o, is less than in the unregulated equilibrium. The optimum can be supported as an equilibrium if travellers are forced to pay a total price of MC^o. Because the price of a trip is the sum of the individual's physical travel cost and the toll, the requisite toll is $\tau^o = MC(q^o) - C(q^o) = q^o \cdot \partial C(q^o)/\partial q$, where $q^o \cdot \partial C(q^o)/\partial q$ is the marginal congestion cost imposed by a traveller on others. This toll is known as a 'Pigouvian' tax, after Pigou (1920).

The fact that a toll is required to support the social optimum reveals the fallacy that travellers fully pay for the congestion they cause through the time they personally lose. Analogously, a person squeezing through a crowd, a shopper queuing in line at a supermarket counter, or a person reading a popular library book after a long wait for it, impose costs on others that they do not themselves bear. Note, however, that at the optimum as drawn in Figure

1, some congestion remains since the generalized cost net of the toll, C^o, exceeds the free-flow cost, C^{ff}. Efficient tolling therefore does not necessarily dictate that congestion be eliminated.

The efficiency gain derived from the optimal toll can be expressed as the increase in social surplus, defined by the reduction in total costs minus the reduction in total benefits due to the decrease in traffic. This gain is measured by the shaded area in Figure 1. The toll revenue, $q^o \cdot \tau^o$, nets out because it is a transfer from road users to the government. Nevertheless, the transfer leaves road users worse off. The q^o users who continue to travel per unit time suffer a cost increase of $MC^o - C^n$. And the $q^n - q^o$ individuals who stop travelling suffer a loss of surplus that ranges from zero for the pre-toll marginal user at q^n, to $MC^o - C^n$ for the new marginal user at q^o. These losses are the root of opposition to congestion pricing. As discussed in Section 6, ways are being sought of using toll revenues to make congestion pricing more politically palatable.

3. Congestion pricing in time-dependent models

Time-dependent models of congestion build on time-independent models by adding two elements: a specification of how travel demand depends on time, and a specification of how traffic flows evolve over time and space. To maintain focus on the time elements of congestion pricing, attention will be limited as in Section 2 to a single road joining one origin and one destination. Until near the end of Section 3 it will also be assumed that the total number of trips is fixed; *i.e.* price inelastic. But heterogeneity in the trip-timing preferences and time costs of travellers is allowed.

First consider the modelling of demand. In Vickrey's (1969) pioneering approach, an individual, i, is assumed to have a preferred time, t_i^*, to complete a trip, and to incur a *schedule delay cost* $D_i(t - t_i^*)$ if arriving at time t instead, where $D_i(0) = 0$ and $D_i(x) \geq 0$ for $x \neq 0$. Let $T(t)$ denote travel time or trip duration, α_i denote i's unit cost of travel time, and $\tau(t)$ denote the toll (if any) at time t. The cost incurred by i in arriving at time t is assumed to be linear in trip duration and additively separable:

$$C_i(t) = \alpha_i T(t) + D_i(t - t_i^*) + \tau(t). \qquad (1)$$

Time-dependent models also specify how speed and flow evolve over time and space. Various modelling approaches have been developed; see Hall (1999) and Lindsey and Verhoef (2000). It is assumed here that traffic flow is governed by a form of flow congestion with no overtaking possible, consistent with what is assumed for the steady state in Figure 1. Details of how flow changes with time and location on the road are unnecessary for purposes here and are therefore omitted. Vickrey's bottleneck model provides an alternative description of supply in which congestion takes the form of queuing. The bottleneck model yields distinctive results in terms of congestion pricing that will be noted in due course.

This section focuses on first-best congestion pricing, leaving second-best pricing considerations to Section 4. The analysis proceeds by first characterizing the no-toll equilibrium (NTE), then the social optimum (SO), and finally identifying the effects of congestion pricing by comparing the NTE and SO.

The NTE is a Nash equilibrium in which each individual minimizes her trip cost defined by (1) with $\tau(t) = 0$, while taking the travel time choices of other travellers as given. Individual *i*'s choice of *t* can be characterized using indifference curves by fixing trip cost parametrically at C_i and solving (1) with $\tau(t) = 0$ for the implied travel time, denoted $T_i(t, C_i)$:

$$T_i(t, C_i) = (C_i - D_i(t - t_i^*)) / \alpha_i . \quad (2)$$

$T_i(t, C_i)$ can be interpreted as a *congestion delay indifference curve*. The absolute slope of $T_i(t, C_i)$, $\left\| D_i'(t - t_i^*) \right\| / \alpha_i$, reflects *i*'s marginal willingness to incur delay in order to arrive closer to t_i^*. The curve is steep if either marginal schedule delay cost is high or unit travel time cost is low. An individual with a steep curve can be said to have a high *congestion tolerance.*

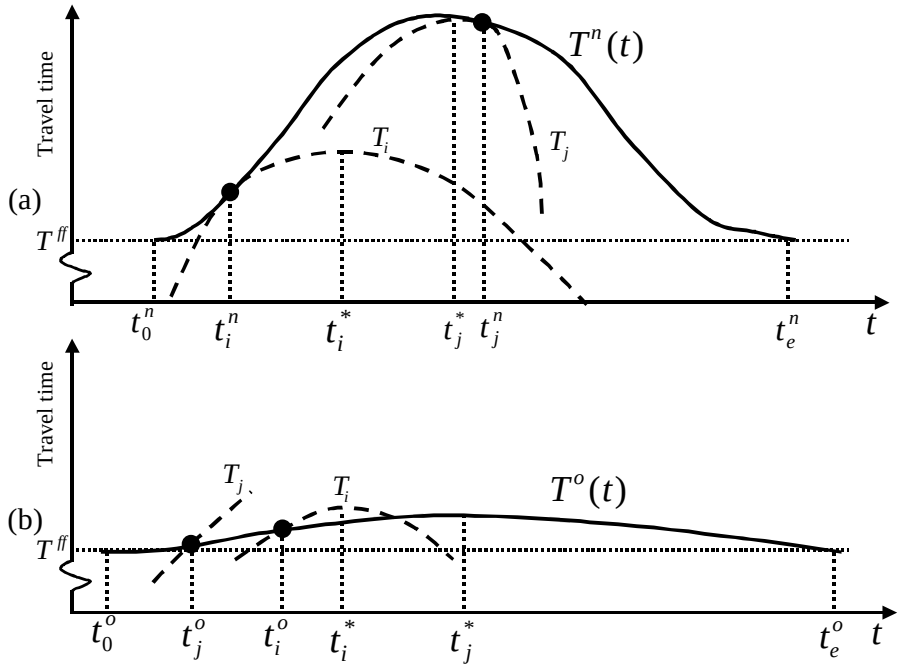

Figure 2. No-toll equilibrium (a) and social optimum (b)

Figure 2(a) depicts a NTE for a time span encompassing a morning peak period. The curve $T^n(t)$ shows how travel time starts rising above free-flow travel time (T^{ff}) at time t_0^n, grows smoothly to a maximum, and then falls back to free-flow conditions at t_e^n. Choice of trip time is shown for two individuals, *i* and *j*. Individual *i* is a highly-paid professional with an early preferred work start time, t_i^*, a very high value of time (α_i), and a strong but somewhat weaker preference for arriving on time — so that *i* has a relatively low congestion tolerance. Individual *i* chooses to arrive early at time $t_i^n < t_i^*$ where the travel time curve $T^n(t)$ is tangent to the dashed curve T_i: the lowest attainable of *i*'s congestion delay

indifference curves. By choosing to arrive early, *i* gains a saving in travel time that more than compensates for the schedule delay. Individual *j* is a service-sector employee with an official work start time t_j^* near the peak, a relatively low value of time (α_j), and little aversion to arriving early, but a strong aversion to arriving late because of the risk of employer censure. Individual *j* chooses to arrive late at $t_j^n > t_j^*$ when travel time is declining. But because *j* has a high congestion tolerance for avoiding late arrival, *j* is unwilling to arrive late by very much.

In the NTE, the travel time curve $T^n(t)$ forms an upper envelope of all the travellers' equilibrium indifference curves, its slope at each point matching the congestion tolerance of the traveller who arrives then. For reference in Section 4, note that heterogeneity in preferred arrival times has a moderating effect on congestion: when some individuals want to arrive early and others late, there is less competition for road space than if everyone wants to arrive at the same time. Heterogeneity in congestion tolerance also has a moderating influence: individuals with a low tolerance travel at the beginning and at the end of the peak period, so contribute relatively little to the buildup of congestion.

Now consider the social optimum (SO). The SO is derived by choosing the time at which each individual arrives to minimize the sum of aggregate schedule delay costs and aggregate travel time costs. This is an optimal control problem that involves both a choice of the *rate* at which individuals arrive and a choice of the *order* in which they arrive. (The relationship between the arrival rate and the departure rate from the origin is determined by the flow congestion curve. Given no overtaking, the order of departure and the order of arrival coincide.) The optimal arrival rate is governed by a tradeoff between schedule delay costs and travel time costs. A high arrival rate compresses the arrival period and reduces schedule delay costs, but boosts congestion and travel delay. Correspondingly, a low arrival rate increases schedule delay costs but dampens congestion. Let $\left(t_0^o, t_e^o\right)$ denote the arrival period. The arrival rate at t_0^o is restricted to maintain free-flow speed, since otherwise it would be better to let the first individual arrive a moment earlier to prevent congestion with only a minimal increase in schedule delay. Similarly, the arrival rate at t_e^o is restrained to free-flow conditions. In between t_0^o and t_e^o, speeds are below the free-flow level. But travel time costs are still reduced relative to the NTE. For this reason the order of arrival in the SO is governed primarily by the goal of minimizing schedule delay costs, which dictates that individuals with strong arrival time preferences arrive at, or near to, their preferred times.

Figure 2(b) shows a SO that might obtain with the same set of travellers as in the NTE of Figure 2(a). (The indifference curves T_i and T_j in Figure 2(b) are explained later.) Arrivals occur over a longer time interval than in the NTE, and the travel time curve, $T^o(t)$, is lower and flatter than $T^n(t)$. Individual *i* still arrives early, at t_i^o. But since *i* has strong arrival time preferences, *i* arrives close to t_i^*. Individual *j* on the other hand has only a weak aversion to arriving early. So *j* is scheduled to arrive early in the rush hour and *before i*, rather than after

t_j^* as in the NTE. Thus, the SO not only involves changes in individual arrival times relative to the NTE, but can also feature changes in their arrival order.

Properties of the SO differ somewhat in the bottleneck queuing model, in which travel speed remains at free-flow speed for flows right up to the bottleneck's capacity. Flow is maintained at capacity throughout the travel period to maximize throughput while avoiding any queuing. Therefore there is no tradeoff between schedule delay costs and travel time costs. Because the arrival rate matches bottleneck capacity in both the SO and the NTE, the duration of the travel period is the same, although arrivals can begin either earlier or later in the SO than in the NTE.

With descriptions of the NTE and SO in hand it is now possible to consider congestion pricing. The question naturally arises whether the SO can be decentralized by tolling. Arnott and Kraus (1998) show that it indeed can be as long as travellers cannot overtake each other, and provided the toll can be varied freely over time. This is a deep result that takes time to appreciate fully. It follows essentially from the fact that the only choice drivers are assumed to have is when to arrive, and the externalities associated with arrival time can be fully internalized through a time-varying toll.

For ease of reference the optimal time-varying congestion toll will henceforth be called the *fine toll*. The fine toll incorporates both a static component analogous to the Pigouvian tax in Section 2, and a dynamic component; see Arnott and Kraus (1998, eqn. 20). (Carey and Srinivasan (1993) derive an equivalent toll for a model with exogenous trip-timing. Yang and Huang (1997) also derive the toll for a variant of the bottleneck model in which the bottleneck's capacity depends on the length of the queue behind it.) Because the fine toll depends not only on the flow congestion technology, but also on the joint frequency distribution in the population of preferred arrival times, values of time, and schedule delay costs, its time evolution can be quite complex. And the toll does not necessarily rise and fall in perfect synchrony with contemporaneous congestion (Carey and Srinivasan, 1993). But because free-flow conditions are maintained at t_0^o and t_e^o, the toll must be the same at these times, as well as before and after the peak period when there is no congestion. With price-inelastic demand any constant amount can be added to or subtracted from the fine toll throughout the day without upsetting the SO.

Reconsider the SO shown in Figure 2(b). Individual *i*'s congestion delay indifference curve, T_i, is steeper at t_i^o than the travel time curve $T^o(t)$. To induce *i* to arrive at t_i^o, rather than earlier or later, the fine toll must increase at the appropriate rate. Similarly, the toll must increase at the appropriate rate to induce individual *j* to arrive at t_j^o. Later in the travel period, the toll has to fall in order to return at t_e^o to the level at t_0^o.

How is the fine toll affected if demand is price-elastic? In the static model of Section 2 in which demand is elastic, the optimal congestion toll is positive in order to deter individuals from taking socially unwarranted trips. Thus, in the dynamic model too it would appear to be

necessary to add a positive constant to the fine toll in order to restrict demand. Yet Arnott and Kraus (1998, Section 4.2) show that this is not the case (see also Carey and Srinivasan, 1993, Section 3). The fine toll supports not only an optimal time pattern of trips conditional on a given demand, but also the optimal set of users. (As discussed in Section 4, this is not true when the time variation of tolling is constrained.) To see this recall that there is no congestion at the beginning or end of the travel period so that individuals who arrive then (or outside the peak period) do not create a congestion externality and therefore should pay no toll.

Consider now the efficiency gains from the fine toll. The toll brings about a reduction in travel time costs — which in the bottleneck model amounts to the full cost of queuing in the NTE. The toll has opposing effects on aggregate schedule delay costs: by spreading out the travel period it tends to boost costs, but by reordering individual arrivals according to strength of travel time preference it reduces costs. The net effect on schedule delay costs can go either way, so that the efficiency gains from tolling can be greater or smaller than the savings in travel time costs. (In the bottleneck model there is no spreading of the travel period, so that schedule delay costs either remain unchanged or fall, and efficiency gains equal or exceed the savings in travel time costs.)

Tolling of course affects the welfare of travellers. In the simple world of Section 2 tolling leaves travellers unequivocally worse off. The case is not as clear cut in the dynamic model because of the efficiency gains derived from altering trip timing. Indeed, depending on the congestion technology and the joint frequency distribution in the population of preferred arrival times, values of time, and schedule delay costs, total private costs can rise or fall. On balance, individuals with high values of time (high α) stand to gain more (or to lose less). In Figure 2 for example, individual *i* with the very high value of time enjoys both a reduction in travel time and a decrease in schedule delay in return for paying the toll.

Several policy lessons can now be drawn. First, congestion pricing not only reduces travel times but also affects schedule delay costs. A cost-benefit analysis that considered only travel times could be biased either for or against a congestion pricing project, and might lead either to unwarranted acceptance or rejection of it. Second, the efficiency gains from congestion pricing are of the same order of magnitude as toll revenue, and can even exceed it. By contrast, the efficiency gains from flat tolls computed using static models can be dwarfed by the toll transfers involved. This suggests that the economics of dynamic congestion pricing schemes are not as sensitive to the costs of infrastructure and operation as are the economics of tolling schemes in static models. Third, congestion pricing has welfare distributional effects on travellers that tend to favour those with high values of time. Because value of time is positively correlated with income, this is consistent with the conventional view that tolling is regressive. Finally, under the fine toll all individuals pay the full marginal social costs of their trips, regardless of their respective characteristics and of when they travel. Section 4

following extends consideration to various real-world complications to examine the robustness of this auspicious result.

4. Second-best issues in congestion pricing

Sections 2 and 3 have outlined the principles of congestion pricing when tolls can be set to match the external costs generated by each traveller. Such pricing is called 'first-best' congestion pricing because it supports a first-best optimum in which roads are used at maximum efficiency. Although useful as a theoretical benchmark, first-best pricing is increasingly recognized as of limited practical relevance. Attention has turned in the recent literature to more realistic types of 'second-best' congestion pricing, in which various costs or constraints deter or prevent the setting of first-best tolls. Examples of second-best tolling include the use of toll-cordons around cities instead of tolling each road in the network, the use of step-tolls instead of smoothly time-varying tolls, tolling according to a fixed daily schedule rather than day-specific traffic conditions, *etc*. The rules for setting optimal second-best tolls are generally quite complicated because they must reflect all sorts of indirect effects, both good and bad. (For an overview of second-best pricing see Bohm, 1987.) This section will discuss a number of examples of second-best congestion pricing without attempting a general treatment.

4.1. Networks

The first-best rules for tolling a single road were identified in Sections 2 and 3. As Beckmann *et al*. (1956), Dafermos (1973) and Yang and Huang (1998) have shown for the static modeling framework, these rules continue to apply to each link of a road network provided every link is efficiently priced. But for several reasons it is quite unlikely that tolling will be implemented throughout a network. First, collecting tolls is costly. Conventional toll-booths involve substantial investments in space-intensive infrastructure, have high operating costs, and delay travellers when they stop to pay. Electronic tolling has much lower operating costs and imposes no delay, but it too requires investments in road-side infrastructure as well as a means of vehicle identification. Because of these collection costs it is generally not economic to toll every street, particularly in a large urban road network. A second constraint on road pricing is that most countries where pricing has been implemented require that toll-free alternatives to toll roads exist. And third, due to the expenses and political resistance to road pricing, pricing is likely to be implemented incrementally rather than all at once. The US for example now has a few demonstration projects that feature 'value pricing', whereby one of several parallel highway lanes is tolled while the other lanes remain free. Thus, even under optimistic assessments about the future of road pricing, much of the road network is liable to remain untolled for a long time. This raises the question how second-best tolls should be set on toll roads given unpriced congestion on untolled roads elsewhere in the network.

Lévy-Lambert (1968) and Marchand (1968) were the first to address this question using a simple network featuring one toll road (call it road *A*) and one untolled road (road *B*) running in parallel between a common origin and destination. With no toll, *B* is overutilized. Excessive usage of *B* can be alleviated by reducing the toll on *A* below its first-best level in order to draw traffic off *B* and on to *A*. The optimal second-best toll is determined by balancing the gains from reducing usage of *B* against the costs of inducing excessive usage of *A*. Verhoef, Nijkamp and Rietveld (1996) demonstrate that if route *B* is particularly congestion-prone, the optimal second-best toll on *A* can be negative. More generally they show that the optimal toll depends on the relative free-flow travel times and capacities of the two routes, and on the price elasticity of travel demand. They also find that the welfare-gains from second-best pricing are typically only a small fraction (*e.g.* 10%) of the gains from first-best pricing. Liu and McDonald (1998) corroborate these results using model parameters descriptive of one of the California road pricing demonstration projects (State Route 91 in Orange County).

These studies may underestimate the efficiency gains from second-best tolling because they ignore some of the ways in which pricing can alter driver behaviour. Braid (1996) and De Palma and Lindsey (2000) allow for trip-timing adjustments by considering time-varying tolls in the Vickrey (1969) bottleneck model applied to the same two-parallel-routes network. Second-best tolling yields higher absolute efficiency gains than in the static model, as well as a greater fraction of the first-best efficiency gains. This is because the toll not only curbs excessive total usage, but also eliminates queuing on the tolled route.

Another way in which second-best tolling can enhance efficiency is through the sorting of drivers according to value of travel time. A pay-lane for example offers an expensive but quick trip that attracts users with high values of time, while the untolled lanes offer cheaper but slower service that caters to other travellers. Using a model with two groups of travellers with different values of time, Small and Yan (1999) find that the efficiency of pay-lanes relative to the first-best optimum is higher than in the equivalent model with no heterogeneity in value of time. Verhoef and Small (1999) obtain broadly similar results using a frequency distribution of value of time based on a Dutch survey of morning peak road users.

The two-parallel-routes network is just one of many network settings where second-best congestion pricing is relevant. A similar setting arises when travellers have a choice between driving and using public transit (Tabuchi, 1993). Public transit systems typically feature significant economies with respect to ridership that often outweigh any congestion externalities. The marginal social cost of a passenger trip is then below average cost, and first-best marginal-cost pricing results in a deficit. If transit is obliged to be self-financing for political or other reasons, then fares must be set at average cost and overpricing of transit results. The second-best toll on the road is then set above marginal cost too in order to boost transit ridership. If, alternatively, the self-financing constraint applies to transportation as a

whole, then the road toll should again be set above marginal cost to cover part of the transit deficit and allow transit fares to be set closer to marginal cost.

Another instance of second-best pricing arises in the setting of public parking fees when roads are underpriced. Glazer and Niskanen (1992) consider a network in which users of public parking, users of private parking and through traffic all drive on the same road link. Glazer and Niskanen derive the optimal combination of lump-sum parking fees, hourly parking fees and capacity for public parking. They show that if an optimal road toll can also be set, then the optimal lump-sum parking fee is zero, and the optimal hourly fee equals the marginal cost of supplying parking space per hour. In the second-best solution in which no road toll is levied, the first-best optimal hourly parking rate still applies, whereas a positive lump-sum parking fee is levied equal to a fraction of the first-best road toll. The lump-sum fee does not fully substitute for the road toll because, by suppressing trips by public parking users, it exacerbates excessive travel by private parking users and through traffic. Thus, analogous to the two-parallel-routes setting, the second-best congestion toll (*i.e.* the lump-sum part of the parking fee) is set below the first-best congestion toll , although the second-best tax rules differ between the two cases.

A general treatment of second-best congestion pricing in a network is found in Verhoef (1998), who derives optimal static tolls on any subset of links (including parking spaces) in an arbitrary network. The toll formulae, which are quite complicated, include terms reflecting marginal external costs on other links, and weights that depend on various demand and cost elasticities. Because the information required to use these formulae may be very costly to obtain, third-best pricing (*i.e.* setting 'quasi' first-best tolls, as if second-best distortions do not exist) or other rules of thumb may be worth considering in practice. Careful consideration, however, is in order in such cases: the use of more or less arbitrary tolls, on a few links of a network only, may well lead to a welfare reduction compared to the no-toll situation.

Closely related to the task of setting second-best tolls is the task of deciding which links in a network to toll, also considered in Verhoef (1998). Some locations where road pricing has been adopted or tested, including Singapore, Hong Kong and Bergen, have well-defined geographical boundaries that simplify the decision. But in many areas this is not the case, and designers must contend with the problem that (as in the two-parallel-routes network) congestion gets displaced from tolled roads to untolled roads. The best candidates for road pricing appear to be freeways and major urban arterial roads because of their high traffic volumes, their role in providing rapid travel over longer distances, and the fact that they do not have close substitutes.

Evidence that congestion pricing may well produce adverse network effects is suggested by a recent study by May and Milne (2000). Using steady-state equilibrium simulations on a road network in Cambridge, UK, they compare congestion pricing with three other road pricing schemes: cordon pricing, time-based pricing, and distance-based pricing. The tolls

considered were not second-best optimal tolls as described above, but various exogenously determined toll levels instead. All four schemes were found to be prone to adverse boundary effects, including "rat running" (where road users seek untolled routes). More discouragingly, a smaller percentage of travellers enjoyed travel time reductions with congestion pricing than with the other schemes. By inducing travellers to re-route to less congested roads, congestion pricing also had a tendency to increase travel distances — with potentially adverse environmental effects. These findings suggest that route-choice decisions deserve particular scrutiny in the design and evaluation of real-world congestion pricing projects. De Borger & Proost (2000) report on the relative efficiency of km-charges, fuel charges, parking charges and public transport pricing for different cities and countries, using the multi-modal TRENEN-model. They find that parking charges combined with cordons can achieve efficiencies of more than 70% of the first-best ideal in some cases, while the potential of fuel pricing and public transport pricing is rather limited.

4.2. Heterogeneity of users

Both road vehicles and travellers vary in a number of characteristics. Vehicles differ in the road space they occupy, the visual obstruction they impose on drivers of other vehicles, their weight and acceleration capabilities, and the number of people they carry. Travellers differ in their values of time, trip-timing preferences, desired speed and so on. Important questions in the practical design of congestion pricing schemes are whether first-best congestion pricing can still be implemented given these dimensions of heterogeneity, and if it cannot be how second-best optimal tolls are determined. In addressing these questions it is useful to distinguish between tolling schemes that are constrained to be *anonymous*; *i.e*. independent of driver or vehicle type, and schemes that can impose *nonanonymous* or type-specific tolls.

Consider first heterogeneity in drivers' values of time and trip-timing preferences. As mentioned in Section 3, Arnott and Kraus (1998) have shown that first-best pricing remains possible using anonymous tolls with heterogeneous user groups provided the tolls can be varied freely over time. The optimality of anonymous tolling derives from the fact that the appropriate toll depends only on the marginal externality costs that drivers impose, and not on their individual preferences.

Optimal anonymous tolling may entail segregation of vehicle or driver types onto separate routes or traffic lanes. Using a static model Verhoef and Small (1999) consider differentiation of tolls across parallel traffic lanes. The higher priced lanes attract drivers with high values of time who are willing to pay for quicker trips, leaving other drivers to use the cheaper but slower lanes. For each lane separately an anonymous toll is still optimal, and efficient segregation of drivers is achieved without other forms of regulation. Nevertheless, the extra benefits turn out to be rather small, so that a second-best single toll applied to the entire highway does not impose much of a welfare loss.

The benefits from having drivers with different values of travel time use different routes must be weighed against the benefits from having drivers with different trip-timing preferences travel on the same roads at different times (see Section 3). Arnott *et al*. (1992) investigate this tradeoff using the bottleneck model with two driver groups and two routes when pricing is limited to time-invariant tolls. Spatial segregation of the groups onto separate routes turns out to be desirable when the groups have a similar congestion tolerance so that little is gained from segregating them temporally. Spatial segregation can be accomplished by imposing differential anonymous tolls on the two routes. A similar analysis could be done for time-varying tolls, which would produce a different set of conditions for optimal segregation.

Next, consider heterogeneity in travel speed — which may be due to differences in driver preferences or vehicle capabilities. Verhoef *et al*. (1999) investigate optimal pricing for two groups using a road on which overtaking is impossible. Traffic is assumed to be sufficiently light that in the absence of desired speed differences there would be no congestion and no need for congestion tolls. Verhoef *et al*. show that the optimal toll for slow vehicles is always higher than the optimal toll for fast vehicles, and *decreases* with the fraction of slow vehicles in total traffic. The toll decreases for two reasons: first a slow vehicle delays fewer fast vehicles on average, and second the average speed of fast vehicles declines asymptotically toward the speed of slow vehicles so that addition of a slow vehicle delays a given fast vehicle by a lesser amount.

Implementation of the optimal pricing scheme requires nonanonymous tolls because slow vehicles must pay more than fast vehicles. This is straightforward if vehicles are observationally distinguishable, such as cars and trucks. It may also be feasible for different groups of car drivers by measuring speeds and using automatic vehicle identification technology. But for practical or political reasons congestion tolls may be constrained to be equal. Verhoef *et al*. (1999) show that the optimal second-best toll is a weighted average of the congestion externalities created by each group. The formula for the toll is a special case of the general formula for second-best tolls derived in Verhoef (1998).

There are obvious advantages to segregating vehicles with different travel speeds onto separate lanes or routes. In practice it is done in rough-and-ready fashion on multilane roads by encouraging or requiring slow-moving vehicles to use the shoulder lanes, as well as by prohibiting slow-moving vehicles from certain routes. In principle, it could be accomplished by tolls. Tolls might also be based on other aspects of driving style that affect congestion and safety (*e.g*. nonuse or misuse of headlights, failure to maintain steady speeds on mountain roads, lane weaving, parking on highway shoulders) using sophisticated electronic monitoring. Such monitoring would supplement or replace current methods using police surveillance, fines and demerits for driving infractions, and adjustments of insurance premiums.

Heterogeneity in the physical characteristics of vehicles can be treated along similar lines to heterogeneity in drivers. Indeed, the principles of static congestion pricing with heterogeneous vehicle types were established by Dafermos (1973). Charging by vehicle type requires nonanonymous tolling. Currently this is done according to vehicle class (car, bus, truck), number of axles, number of trailers, laden weight and axle weight. Traffic engineers often account for the greater contribution of trucks and buses to congestion by computing passenger car equivalents that depend on type of road and terrain, and tolls could be set on this basis.

4.3. Step tolls

The discussion of dynamic congestion pricing in Section 3 focused exclusively on the optimal continuously time-varying (fine) toll. Fine tolling is certainly feasible using electronic toll collection technology. But there is conflicting evidence on whether drivers appreciate smoothly-varying tolls. And most existing road pricing schemes employ either constant tolls, or step (piecewise constant) tolls — as is the case with Singapore's former area licensing scheme and Trondheim's toll ring.

Because step tolls increase or decrease in jumps, they do not rise or fall in perfect synchrony with congestion externalities and are therefore not fully efficient. Indeed, surges of traffic can occur just before increments and just after decrements, as were observed in Singapore and (to a lesser extent) in Trondheim. Arnott *et al*. (1993) show that in the bottleneck model with inelastic demand a one-step toll provides one half or slightly more of the efficiency gains of the fine toll. Chu (1999) obtains broadly similar results in a more elaborate model featuring flow congestion and a traveller population with heterogeneous values of time and schedule delay costs. The efficiency of a step toll scheme naturally improves with the number of steps it embodies (Laih, 1994). But unlike with a fine toll, marginal-cost pricing requires that a fixed component be added to the step toll in order to support the optimal number of trips under elastic demand. Thus, the optimal step toll is positive at the beginning and end of the congested travel period.

As noted above, anonymous tolling is efficient with heterogeneous travellers when tolls can be varied freely over time. Since this is not the case of step tolls, nonanonymous tolling can in principle enhance the benefits of a step-toll congestion pricing scheme, although setting tolls according to values of time and time-of-use preferences is unlikely to be possible at a high level of precision.

4.4. Uncertainty and information provision

To this point it has been assumed that travellers know how long a trip will take and what cost they will incur, inclusive of toll. In reality travellers are usually faced with one or both of two types of uncertainty: idiosyncratic uncertainty and objective uncertainty. *Idiosyncratic uncertainty* exists when traffic conditions are predictable, but individual travellers do not

know travel times precisely and form their own idiosyncratic perceptions. The standard approach to describing traveller behaviour given idiosyncratic uncertainty is as a Stochastic User Equilibrium (Daganzo and Sheffi, 1977) in which drivers minimize their perceived travel costs. Smith *et al.* (1995) and Yang (1999) have shown that under reasonable assumptions standard static Pigouvian tolls based on actual travel costs remain optimal in the Stochastic User Equilibrium framework.

In contrast to idiosyncratic uncertainty, *objective uncertainty* exists when travel conditions vary unpredictably on account of traffic accidents, bad weather, roadwork, surges in demand due to special events or transit strikes, and so on. Various studies (e.g. Schrank and Lomax, 1999) have determined that a large fraction of time lost in congestion is attributable to these shocks. One way to model traveller behaviour in the face of objective uncertainty parallels the treatment of idiosyncratic uncertainty in Stochastic User Equilibrium. In this approach, called Stochastic Network Stochastic User Equilibrium by Emmerink (1998, Ch. 4), drivers minimize their expected trip costs conditional on any information about traffic conditions available to them from radio or other sources.

Two approaches to congestion pricing with objective uncertainty are possible. One is to charge tolls that do not vary with traffic conditions. For example, the toll for trucks on Route *A* at 8 a.m. on weekdays would be the same every day regardless of how congested Route *A* actually is. The other approach, termed *responsive* pricing by Vickrey (1971), is to condition tolls on any information available about traffic conditions in order to match actual congestion externality costs as closely as possible. To implement responsive pricing effectively it is necessary to collect information about traffic conditions, calculate the appropriate tolls, and convey the updated information and tolls to drivers — all on an ongoing basis. While infeasible in the past, this goal is becoming practicable through the use of Advanced Traveller Information Systems (ATIS) that transmit information to travellers by phone or Internet, or directly to vehicles equipped with on-board computers and Global Positioning Systems receivers. (For overviews of ATIS see Emmerink *et al.*, 1994, and Emmerink and Nijkamp, 1999.)

Recent research on congestion pricing under uncertainty has focused on the use of either responsive or ‘nonresponsive’ pricing in conjunction with the implementation of ATIS. It is clear that there are technological synergies between ATIS and road pricing as far as their use of road infrastructure, centralized computing capability and communications with drivers. It is not so evident whether there are also synergies in their benefits; that is whether the benefits are superadditive or subadditive. In favour of superadditivity it can be observed that in the presence of unpriced congestion providing information to drivers can be welfare-reducing, as has been shown by Ben-Akiva *et al.* (1991) and Catoni and Pallottino (1991) *inter alios*. In favour of subadditivity it can be argued that if one technology improves travel conditions, it reduces the scope for further gains from the other technology.

A few recent studies have investigated the additivity question. Verhoef, Emmerink, Nijkamp and Rietveld (1996) use a static model featuring endogenous route choice and elastic demand. They find that nonresponsive pricing and (perfect) information are approximately additive in their benefits, as well as complementary in the sense that under conditions when one instrument does not yield much benefit the other instrument does particularly well. El Sanhouri and Bernstein (1994) adopt a dynamic model with endogenous trip-timing decisions, and likewise find the benefits of nonresponsive pricing and ATIS are approximately additive. Yang (1999) finds that additivity depends on how many drivers receive information, while De Palma and Lindsey (1998) show that information can be welfare-reducing unless it is supplemented with *responsive* pricing.

One consideration not addressed in these studies is the attitude of drivers toward tolling under uncertainty. It is unclear as yet from the limited formal research whether drivers are better off in terms of expected travel costs under nonresponsive or responsive pricing. Some surveys have found that drivers dislike uncertainty about how much they will have to pay in tolls. Aversion to uncertainty about payment was one of the reasons for opposition to the congestion metering project planned for Cambridge, England, in which vehicles would have been charged on the basis of actual congestion experienced. However, drivers have been receptive to the recent adoption of responsive pricing on Interstate I-15 north of San Diego. Further research is clearly called for on the economics and politics of road pricing under uncertainty.

4.5. Interactions with other economic sectors

Economic transport models typically consider only demand and supply conditions pertaining to transport, and thus implicitly assume that the rest of the economy operates under first-best conditions. Although this assumption simplifies the modelling, it is usually well off the mark. In particular, distortionary taxes in labour and commodity markets are the norm, motivated by the need to raise government funds. This is recognized in the literature in environmental economics on the controversial 'double-dividend hypothesis', which examines the interactions between environmental externality charges and distortionary taxes (*e.g.* Bovenberg and Goulder, 1996).

Following the environmental economics lead, several recent studies have addressed second-best aspects of road pricing that arise from interactions with the rest of the economy. For a number of reasons a congestion tax on transport is likely to have a non-marginal impact on efficiency elsewhere in the economy. First, the extra taxes on transport as a consumption good reduce the real purchasing power of the wage, thereby aggravating the distortion from a pre-existing tax on labour ("the tax-interaction effect"). Second, by affecting the costs of commuting, peak-period congestion tolls also more directly influence labour supply (the "complementarity effect"). Third, the revenues from congestion tolls can be used in various ways: to finance road capacity expansion or public transit services, to reduce labour taxes, to

increase government spending on other services, and so on (the "recycling effect"). The benefits from these alternative expenditures may vary considerably with the direction and magnitude of the distortions respectively involved. Fourth, all road users pay the tax but not all road users will share in the tax revenues ("the tax shifting effect"). This may lead to direct welfare effects if different groups have different welfare weights in the social welfare function, and may furthermore induce efficiency effects through interactions with the other efficiency effects mentioned.

Mayeres and Proost (1999) evaluate the efficiency effects of transportation charges in Belgium by computing the marginal welfare cost of public funds for a number of tax instruments using a general equilibrium model, for different degrees of social income inequality aversion. They find that a marginal increase in peak-period road transport prices yields the highest benefit when revenue is spent on road capacity expansion. The only negative benefit obtains for expenditure on public transport (which is already heavily subsidized) provided the degree of social inequality aversion is not too high, so that the benefits received by lower-income groups are not weighted too heavily.

Parry and Bento (1999) also find that the general equilibrium effects of road congestion pricing schemes are sensitive to the allocation of revenues, and may deviate considerably from partial equilibrium estimates. In particular, they find that a lump-sum redistribution of congestion tax revenues can make the tax welfare-reducing because of its depressing effect on labour supply, whereas using the revenues to lower labour taxes doubles the overall gains. Their results thus demonstrate that incorporation of general equilibrium effects may greatly magnify the difference between welfare when revenues are, or are not, used to reduce pre-existing distortions.

General equilibrium studies such as these provide at least two important lessons. One lesson is that partial equilibrium analyses of road congestion pricing can miss out on important indirect efficiency and welfare effects. Second-best optimal congestion taxes that take these effects into account can differ significantly from the first-best taxes derived using partial equilibrium methods. The second lesson is that it is dangerous to allocate revenues from congestion pricing solely to 'buy' public acceptability (see also Section 6 below). Poorly chosen allocations can have such adverse indirect effects that they outweigh the direct benefits of congestion pricing, leaving society worse off.

4.6. Other traffic-related externalities

Accidents, noise, local air pollution and global warming are all negative side effects of road traffic. By altering the timing, location, speed and volume of traffic levels, congestion pricing affects these other externalities. But these externalities cannot be properly internalized through pricing aimed solely at congestion relief. While congestion may contribute to extra fuel consumption and pollution emissions, it has proved difficult to establish a tight statistical relationship between them (Small and Gómez-Ibáñez, 1999, Section 3.2). And by speeding up

traffic, congestion pricing may reduce the frequency of accidents but increase their severity, although again statistical evidence is hard to come by. Thus, it is safe to say that the incidence of congestion and the incidence of other transport externalities are far from perfectly correlated. Additional policy instruments are required to address these other externalities, such as the gasoline tax, periodic vehicle emissions testing, road design and safety legislation, *etc*. These other instruments are unlikely to operate perfectly, however, so that the impact of congestion pricing on other externalities should in principle be accounted for.

4.7. Congestion pricing by private operators

Private toll roads have long existed in Europe and the Pacific Rim (Gómez-Ibáñez and Meyer, 1993), and privately operated pay-lanes are emerging in the US, and being considered for the Randstad area in The Netherlands. Private toll-road operators are typically interested in maximizing profit rather than social surplus, so that first-best pricing cannot be expected of them. The purpose of this section is to compare profit-maximizing and welfare-maximizing congestion tolls on the same facilities. To simplify, regulatory constraints on private tolls — which are often imposed in the form of price or rate-of-return caps — are ignored, as are other distortions from first-best conditions.

Consider first monopoly-pricing. Various (static) monopoly settings have been considered in the literature, including a single-road monopolist competing with train service (Edelson, 1971), monopoly pricing of a single facility facing no competition (Mills, 1981; Mohring, 1985) and a monopolist operating two roads (Verhoef, Nijkamp and Rietveld, 1996). As long as all potential users incur the same disutility from congestion, the monopoly toll in each of these settings is found to be

$$\tau_i = N_i \partial C_i / \partial N_i - N_i \partial p / dN \qquad (3)$$

where τ_i is the toll on road *i* (possibly the only road), N_i is usage of road *i*, $C_i(N_i)$ is travel cost on road *i*, *N* is total usage and $p(N)$ is the inverse demand curve for travel. The first component of the toll in eqn. (3) is the first-best congestion toll, and the second component is a markup that depends on demand but not travel costs. The monopolist therefore fully internalizes the congestion externality. The reason is that by reducing congestion costs the monopolist can charge a higher toll without losing patronage. But, as is true of monopolists in general, the monopolist adds a markup that is larger the steeper is demand. (Indeed, by rewriting eqn. (3) as $\tau_i + N_i \partial p / \partial N = N_i \partial C_i / \partial N_i$, it can be seen that the monopolist equates marginal revenue, rather than price, to marginal social cost.) Thus, except in the limiting case of perfectly elastic demand, the monopolist sets a toll above the first-best toll and accommodates too little traffic. Indeed, the markup may be so high that welfare is actually reduced relative to no tolls (*e.g*. Verhoef and Small, 1999).

In most countries (and in California) private road monopolies are not allowed because of legislation requiring that free alternatives be available (Gómez-Ibáñez and Meyer, 1993). Accordingly, several studies have examined private-sector road pricing using the two-parallel-

routes network (discussed under 'Networks' above) where one of the two routes is private, and the other is either free-access or publicly operated. Using static models, Verhoef, Nijkamp and Rietveld (1996) and Liu and McDonald (1998) show that the private operator will again set a toll higher than the second-best toll, and may end up reducing welfare relative to no toll. De Palma and Lindsey (2000) show that the welfare effects of private-sector pricing are more salubrious in a dynamic model with endogenous trip-timing decisions. As in the case of second-best pricing (see 'Networks' above) this is because a private operator has an incentive to adopt time-varying tolls to reduce peak-period congestion.

Whether private toll roads can operate profitably depends on various factors, including infrastructure and operating costs, regulatory constraints on tolls, and competition from other roads and modes of transport (Nijkamp and Rienstra, 1995). Viton (1995) concludes that private toll roads can be profitable under a wide range of conditions even when competing against a free public road, particularly for urban areas in which high tolls can be charged during peak periods. In part, Viton's results are driven by his assumption that drivers have strong idiosyncratic preferences for roads, which limits their willingness to switch from one to the other. He also assumes that a private operator can price discriminate on the basis of automobile size, with the result that in his model the toll per mile for large cars is over triple the toll for small cars.

One consideration ignored in the discussion thus far is heterogeneity in drivers' congestion costs. As mentioned above under 'Heterogeneity of users', welfare-maximizing tolls depend on the congestion costs incurred by all drivers including inframarginal ones. By contrast, as both Edelson (1971) and Mills (1981) observe, (undifferentiated) profit-maximizing tolls depend only on the congestion costs incurred by marginal users. When inframarginal users are more averse to congestion on average than are marginal users, private tolls are biased downward, and in theory might even be lower than first-best tolls. Nevertheless, like a welfare maximizer a private firm operating more than one facility has an incentive to provide differentiated quality service (see Chander and Leruth, 1989) by charging different tolls on different roads or traffic lanes. Because firms are inclined to set high tolls that may nearly eliminate congestion, however, the scope for differentiation on the basis of speeds may be (very) limited (Verhoef and Small, 1999) – although it is still possible when the routes differ in length.

5. Congestion pricing and investment

Building new roads was once almost a conditioned response to road congestion. But it is now severely constrained by environmental concerns, and by shortages of funds and space. Still, selective construction continues, and as roads wear out decisions must be made about rehabilitation and replacement. How much should be spent on roads thus remains an important question. The purpose of this section is to explore the dependence of optimal investment on how road usage is priced. (Another interesting question is the degree to which

optimal investments are self-financing under congestion pricing; for summaries and references see Hau, 1998, Sections 3.7, 3.10 and 3.11, and Lindsey and Verhoef, 2000.)

Optimal capacity of a single road is easily characterized. Following Arnott *et al*. (1998, pp.88-89) let K denote capacity, N the number of trips taken, $C(N,K)$ trip cost net of tolls, $G(N)$ total gross benefits from trips, and $B(N,K) = G(N) - NC(N,K)$ total net benefits. Then the marginal benefit of capacity expansion (gross of construction cost) is

$$\frac{dB}{dK} = \frac{\partial B}{\partial K} + \frac{\partial B}{\partial N}\frac{dN}{dK} = \frac{\partial B}{\partial K} + \frac{\partial B}{\partial N}\frac{dN}{dp}\frac{dp}{dK}$$
$$= -N\frac{\partial C}{\partial K} + \left[\left(\frac{\partial G}{\partial N} - C(N,K) - N\frac{\partial C}{\partial N}\right)\frac{dN}{dp}\frac{dp}{dK}\right] \qquad (4)$$

The first term on the right-hand side ($-N\,\partial C/\partial K$) is the direct benefit of the capacity expansion that would occur with no change in travel behaviour. The second term, in square brackets, is the indirect benefit that derives from an increase in the number of trips due to the reduction in trip price caused by the capacity expansion. With marginal-cost pricing, the envelope theorem applies, and the indirect benefit is zero because the marginal social benefit of the induced increase ($\partial G/\partial N$) in traffic equals the marginal social cost ($C + N\,\partial C/\partial N$). But with underpricing (or no pricing) of road use, the marginal social cost exceeds the marginal social benefit. The indirect benefit is therefore negative and, in the limiting case of no tolling and perfectly elastic demand, completely offsets the direct benefit. To see this, consider a stationary traffic setting and suppose that capacity is increased. This shifts the travel cost curve ($C(q)$ in Figure 1) to the right. With no tolling, equilibrium is established at the new intersection of $C(q)$ with the demand curve. If the demand curve is horizontal, traffic volume increases until trip cost is back to its previous level and the investment yields no benefit.

This reasoning might suggest that with underpriced (or unpriced) congestion, optimal capacity is lower than in the first best because of the negative indirect effect. But because underpricing of congestion results in greater usage, the direct benefit of a capacity expansion is higher than in the first best. The net effect of these opposing forces has been investigated extensively in the literature; Arnott and Yan (2000) provide an insightful review and synthesis. One result is that if a toll is reduced slightly below the first-best level, optimal capacity *rises* because the increase in the positive direct effect dominates the increase in the negative indirect effect. (In eqn. (4), both terms of the direct effect, $-N\cdot\partial C/\partial K$, increase to first-order in the capacity expansion, whereas the term in brackets of the indirect effect $\left(\partial G/\partial N - C(N,K) - N\,\partial C/\partial N\right)$ is zero at the first-best solution.)

Another comparison of potential practical importance is between second-best optimal capacity and the capacity that would be chosen if the indirect effects of expansion associated with underpricing of congestion are ignored. Such practice is dubbed “naïve cost benefit analysis”. Arnott and Yan (2000) show that in the standard static model, naïve CBA leads

unambiguously to overestimation of the benefit from expansion and hence to excessive investment. Indeed, they show using a plausible numerical example that the capacity chosen using naïve CBA can exceed second-best optimal capacity by more than second-best capacity exceeds first-best capacity.

Naïve CBA rules have also been investigated using dynamic models of morning peak travel by Small (1992a) and Henderson (1992). Small (1992a, pp.134-137) adopts the bottleneck queuing model with linear schedule delay costs for arriving early or late, and fixed total demand. The planner who chooses capacity is assumed to err by treating the departure times of travellers as given. This introduces two opposing biases: by ignoring that a capacity expansion induces travellers to reschedule trips toward the peak, the benefits from a reduction in schedule delay are missed, but the savings in travel (queuing) time are overestimated. Small finds that the second effect dominates if unit travel time costs are large relative to unit schedule delay costs. This is because the perceived travel time savings are then large, whereas the overlooked benefits from rescheduling are small.

Henderson's (1992) model differs from Small's in two respects. First, travel is subject to *flow* congestion rather than queuing, with travel time determined by the instantaneous arrival rate of vehicles at work. And second, the planner errs by treating the *arrival* times, rather than departure times, of travellers as given. Henderson concludes that the misguided planner overinvests in capacity regardless of the relative unit costs of travel time and schedule delay. The contrast with Small's conclusion highlights the sensitivity of results to assumptions, and suggests that further research on naïve CBA is warranted.

To conclude this section it is instructive to see how the impact of congestion pricing on optimal capacity is affected by some of the complications considered in Section 4.

Networks

If marginal-cost pricing is adopted throughout a network then all the indirect effects of a capacity expansion net out, and the marginal benefit from expanding a single link is still given by the direct effect in eqn. (4). In the absence of marginal-cost pricing of all links, however, a capacity investment can have outright perverse effects. An extreme case of this is the 'Braess paradox' (Braess, 1968) that occurs when adding a new link to an untolled network increases total travel costs.

Step tolls

Eqn. (4) remains valid for any pricing regime, including step tolls, as long as the appropriate trip cost function, $C(N,K)$, is used. Arnott *et al*. (1993) show that in the bottleneck model with homogeneous users a clear-cut ranking of optimal capacity according to the time-sensitivity of the tolling regime obtains. When the price elasticity of demand is less than unity (as it typically is for peak-period travel) optimal capacity is lower the more refined the tolling

regime. This is so because total trip costs are lower for any given capacity in a more refined tolling regime — so that the marginal direct benefit from expansion is lower, and in addition because with relatively inelastic demand the direct effect of expansion dominates the indirect effect. The opposite ranking obtains if the price elasticity exceeds unity.

Congestion pricing by private operators

The profit-maximizing capacity of a private operator can be derived in an analogous way to the socially optimal capacity. Provided the operator chooses a profit-maximizing toll, the envelope theorem applies and the marginal benefit of capacity is given by the same formula as for the direct effect in eqn. (4): $-N\partial C/\partial K$ (see Small, 1992a, p.141). But because the private operator sets a higher toll, usage of the road is less than under public management. Therefore, as long as $-N\partial C/\partial K$ is an increasing function of N, the private operator underinvests in capacity (De Vany, 1976). This is true whatever (common) form of tolling the private and public operator adopt; *e.g.* flat, single-step, multiple-step or fine.

6. The social and political feasibility of congestion pricing

Despite its economic appeal, road congestion pricing appears to enjoy little support outside academia. The limited social and, consequently, political support for congestion pricing has caused many proposed schemes to be abandoned before implementation, or at least to be postponed (sometimes indefinitely). These include detailed plans for Hong Kong, London, The Randstad, and Stockholm (Small and Gómez-Ibáñez, 1998).

One important reason for opposition to congestion pricing was identified in Section 2: before redistribution of toll revenues, everybody except the taxman appears to be worse off. This results should be qualified on two counts. First, if congestion takes the form of pure queuing and users have identical values of time, a fine toll leaves users equally well off. Second, users with a high value of travel time may benefit from congestion pricing even before revenue recycling; see Section 3, Richardson (1974), and supporting empirical evidence from a Dutch survey by Verhoef *et al.* (1997a). Still, because value of time is positively correlated with income, this implies that road pricing is regressive — which is unlikely to improve the social acceptability of congestion pricing.

Because tolling makes many users worse off, revenue allocation and its impacts on income distribution have been identified as key determinants of the acceptability of congestion pricing. Various allocation schemes have been proposed that leave all major user groups better off. Goodwin (1989) and Small (1992b) both propose schemes that allocate revenues in three ways. In Small's (1992b) scheme revenues are used to reimburse travellers as a group, to offset regressive taxes, and to fund new transportation services (particularly transit). Small finds that each of the six prototypical residents he considers benefits from the scheme. In the survey of Verhoef *et al.* (1997a) road users expressed the following preference (in decreasing order) for the use of toll revenues: investment in new roads, reduction in

vehicle ownership taxes, reduction in fuel taxes, investment in public transport, subsidies for public transport, investment in carpool facilities, general tax cuts, and expansion of other public expenditures. The survey suggests that the use of toll revenue to finance road infrastructure as a substitute for other funding sources is politically attractive. It is also transparent, and efficient from an economy-wide perspective according to the general equilibrium analysis of Mayeres and Proost (1999) discussed in Section 4. Indeed, it is worth repeating here that general equilibrium studies of road pricing strongly suggest that revenue allocation schemes designed solely to improve the public acceptability may induce welfare losses elsewhere in the economy, leading to efficiency losses that may even outweigh the initial improvements. A trade-off between efficiency and acceptability impacts of revenue allocation schemes will generally exist, and should be given careful attention in their design.

Apart from the above considerations, which reflect a more or less rational attitude of road users towards congestion pricing, other reasons for opposing congestion pricing have been identified. From a review of public attitude studies Jones (1998) identifies the following reasons for opposition: (1) drivers find it difficult to accept the idea of being charged for something that they wish to avoid (congestion), and also feel that congestion is not their fault but rather something that is imposed on them by others; (2) road pricing is not needed, either because congestion is not bad enough or because other measures are superior; (3) pricing will not get people out of their cars; (4) the technology will not work; (5) privacy concerns; (6) diversion of traffic outside the charged area; (7) road pricing is just another form of taxation; and (8) perceived unfairness. Similar concerns were recently voiced in The Netherlands when the Dutch Automobile Association (ANWB) successfully launched a large public campaign to prevent the implementation of a full-scale congestion pricing scheme for the Randstad area. The Minister of Transport managed to save part of the original plan (as of this time of writing) only after radically downsizing it to a few toll-points and pay-lanes, and after offering extra money for infrastructure investments to the cities affected.

US experience also suggests that congestion pricing may have to begin at a modest level involving a few demonstration projects such as pay-lanes, and may very well end there too. As discussed in Section 4, however, pay-lanes may yield only a small fraction of the potential welfare gains from congestion pricing. Verhoef and Small (1999) suggest that instead of pay-lanes, highway pricing should involve ‘free-lanes’ whereby all but one lane is tolled. This would have the merit of pricing a larger fraction of capacity while still offering via the free lane a lifeline service for travellers with low values of travel time (see also Frick *et al.,* 1996).

Another way to improve the acceptability of road pricing is to reduce the amount of toll collected in the first place. Daganzo (1995) has proposed a combination of pricing and rationing whereby each traveller is prohibited from driving on some fraction of days. As a variant, Daganzo and Garcia (1998) suggest that a fraction of drivers be exempted from paying the toll on any given day. An analogous idea, proposed by Verhoef *et al.* (1997b), is to

issue free of charge a large number of tradeable smart-card units for use with electronic road pricing. Yet another suggestion by DeCorla-Souza (2000) is to convert some existing freeway lanes to toll lanes, and to use the toll revenues to provide drivers who use the remaining free lanes with credits for future trips on the toll lanes, for parking, or for trips by transit.

According to economists, congestion pricing is the *best* instrument for controlling congestion. Yet the bulk of empirical evidence suggests that, even with cleverly designed tolling mechanisms and toll revenue allocation schemes, congestion pricing is likely to remain difficult to implement. It is no coincidence that the only area-wide congestion pricing scheme currently in operation is in Singapore, where the culture allows the government to implement unpopular policies. Indeed, Frick *et al.* (1996) infer from the series of failed attempts to implement congestion pricing on the San Francisco Bay Bridge that both the public and officials must be convinced that *no* feasible alternative to congestion pricing exists before it will be accepted.

7. Conclusion

Given forecasts of continuing growth in road travel, and the reduced scope for expansion of road infrastructure, traffic congestion is not a problem that will go away soon. Recent advances in electronic vehicle identification and automated charging technologies have made congestion pricing a viable means of combating traffic congestion, rather than just an academic *curiosum*. This paper reviews the economic principles behind congestion pricing, which derive from the benefits of charging travellers for the externalities they create. Attention is paid to various complications that make simple textbook congestion pricing models of limited relevance, and dictate that congestion pricing schemes be studied from the perspective of the theory of the second best.

Despite the economic case for congestion pricing, it has attracted strong social and political opposition, and assorted legal and institutional constraints create further barriers to implementation. It thus seems safe to conclude with the prediction that coming decades will witness an increasing number of attempts to implement congestion pricing, many of which will fail in their early stages, but some of which will succeed if only on a piecemeal basis. The design and evaluation of such schemes will probably require a deeper understanding of the direct and indirect impacts of congestion pricing than is available to date. One can hope, therefore, that this review will be obsolete before too long.

References

Arnott, R., K. Arrow, A.B. Atkinson and J.H. Dreze (eds.) (1994). *Public Economics: Selected Papers by William Vickrey*. Cambridge University Press, Cambridge.

Arnott, R. and M. Kraus (1998). When are anonymous congestion charges consistent with marginal cost pricing? *Journal of Public Economics*, **67**, 45-64.

Arnott, R., A. de Palma and R. Lindsey (1992). Route choice with heterogeneous drivers and group-specific congestion costs. *Regional Science and Urban Economics*, **22**(1), 71-102.

Arnott, R., A. de Palma and R. Lindsey (1993). A structural model of peak- period congestion: A traffic bottleneck with elastic demand. *American Economic Review*, **83**(1), 161-179.

Arnott, R., A. de Palma and R. Lindsey (1998). Recent developments in the bottleneck model. In: *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* (K.J. Button and E.T. Verhoef, eds.), pp. 79-110. Edward Elgar, Cheltenham, UK.

Arnott, R. and A. Yan (2000). The two-mode problem: Second-best pricing and capacity. Working paper, Boston College.

Beckmann, M., C.B. McGuire and C.B. Winsten (1956). *Studies in the Economics of Transportation*. Yale University Press, New Haven.

Ben-Akiva, M., A. de Palma and I. Kaysi (1991). Dynamic network models and driver information systems. *Transportation Research A*, **25A**(5), 251-266.

Bohm, P. (1987). Second best. *The New Palgrave: A Dictionary of Economics Vol. 4*, pp. 280-284. Macmillan, New York.

Bovenberg, A.L. and L.H. Goulder (1996). Optimal environmental taxation in the presence of other taxes: a general equilibrium analysis. *American Economic Review: Papers and Proceedings*, **86**, 985-1000.

Braess, D. (1968). Über ein Paradoxen des Verkehrsplanung. *Unternehmenforschung*, **12**, 258-268.

Braid, R.M. (1996). Peak-load pricing of a transportation route with an unpriced substitute. *Journal of Urban Economics*, **40**, 179-197.

Carey, M. and A. Srinivasan (1993). Externalities, average and marginal costs, and tolls on congested networks with time-varying flows. *Operations Research*, **41**(1), 217-231.

Catoni, S. and S. Pallottino (1991). Traffic equilibrium paradoxes. *Transportation Science*, **25**, 240-244.

Chander, P. and L. Leruth (1989). The optimal product mix for a monopolist in the presence of congestion effects. *International Journal of Industrial Organization*, **7**, 437-449.

Chu, X. (1999). Alternative congestion pricing schedules. *Regional Science and Urban Economics*, **29**, 697-722.

Dafermos, S. (1973). Toll patterns for multiclass-user transportation networks. *Transportation Science*, **7**, 211-223.

Daganzo, C.F. (1995). A Pareto optimum congestion reduction scheme. *Transportation Research B*, **29B**(2), 139-154.

Daganzo, C.F. and R.C. Garcia (2000). A Pareto improving strategy for the time-dependent morning commute problem. *Transportation Science*, **34**(3), 303-311.

Daganzo, C.F. and Y. Sheffi (1977). On stochastic models of traffic assignment. *Transportation Science*, **11**, 253-274.

De Borger, B. and S. Proost (eds.) (2000). *Reforming Transport Pricing in the European Union.* Edward Elgar, Cheltenham (forthcoming).

DeCorla-Souza, P. (2000). Making pricing of currently free highway lanes acceptable to the public. *Transportation Quarterly*, **54**(3), 1-20.

De Palma, A. and R. Lindsey (1998). Information and usage of congestible facilities under different pricing regimes. *Canadian Journal of Economics*, **31**(3), 666-692.

De Palma, A. and R. Lindsey (2000). Private toll roads: competition under various ownership regimes. *Annals of Regional Science*, **34**(1), 13-35.

De Vany, A. (1976). Uncertainty, waiting time, and capacity utilization: A stochastic theory of product quality. *Journal of Political Economy*, **84**(3), 523-541.

Edelson, N.E. (1971). Congestion tolls under monopoly. *American Economic Review*, **61**(5), 872-882.

El Sanhouri, I. and D. Bernstein (1994). Integrating driver information and congestion pricing systems. *Transportation Research Record*, **1450**, 44-50.

Emmerink, R.H.M. (1998). *Information and Pricing in Road Transportation*. Springer Verlag, Berlin.

Emmerink, R.H.M. and P. Nijkamp (eds.) (1999). *Behavioural and Network Impacts of Driver Information Systems*. Ashgate, Aldershot.

Emmerink, R.H.M., P. Nijkamp, P. Rietveld and K.W. Axhausen (1994). The Economics of motorist information systems revisited. *Transport Reviews*, **14**(4), 363-388.

Frick, K.T., S. Heminger and H. Dittmar (1996). Bay Bridge congestion-pricing project: lessons learned to date. *Transportation Research Record*, **1558**, 29-38.

Gómez-Ibáñez, J.A. and J.R. Meyer (1993). *Going Private: The International Experience with Transport Privatization*. The Brookings Institution, Washington, DC.

Gómez-Ibáñez, J.A. and K.A. Small (1994). Road pricing for congestion management: A survey of international practice. *National Cooperative Highway Research Program, Synthesis of Highway Practice* 210, TRB, National Academy Press, Washington, D.C.

Glazer, A. and E. Niskanen (1992). Parking fees and congestion. *Regional Science and Urban Economics*, **22**, 123-132.

Goodwin, P.B. (1989). The rule of three: a possible solution to the political problem of competing objectives for road pricing. *Traffic Engineering and Control*, **30**(10), 495-497.

Hall, W.R. (ed.) (1999). *Handbook of Transportation Science*. International Series in Operations Research and Management, Kluwer Academic Publishers, Norwell, MA.

Hau, T.D. (1998). Congestion pricing and road investment. In: *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* (K.J. Button and E.T. Verhoef, eds.), pp. 39-78. Edward Elgar, Cheltenham, UK.

Henderson, J.V. (1992). Peak shifting and cost-benefit miscalculations. *Regional Science and Urban Economics*, **22**, 103-121.

Jones, P. (1998). Urban road pricing: public acceptability and barriers to implementation. In *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* (K.J. Button and E.T. Verhoef, eds.), pp. 263-284. Edward Elgar: Cheltenham, UK.

Knight, F. (1924). Some fallacies in the interpretation of social costs. *Quarterly Journal of Economics*, **38**(4), 582-606.

Laih, C-H. (1994). Queuing at a bottleneck with single- and multi-step tolls. *Transportation Research A*, **28A**(3), 197-208.

Lévy-Lambert, H. (1968). Tarification des services à qualité variable: application aux péages de circulation. *Econometrica*, **36**(3-4), 564-574.

Lindsey, R. and E.T. Verhoef (2000). Congestion modelling. Forthcoming in: *Handbook of Transport Modelling*, *Vol. 1*. (D.A. Hensher and K.J. Button, eds.), Elsevier Science, Oxford.

Liu, L.N. and J.F. McDonald (1998). Efficient congestion tolls in the presence of unpriced congestion: a peak and off-peak simulation model. *Journal of Urban Economics*, **44**, 352-366.

Marchand, M. (1968). A note on optimal tolls in an imperfect environment. *Econometrica*, **36**(3-4), 575-581.

May, A.D. and D.S. Milne (2000). Effects of alternative road pricing systems on network performance. *Transportation Research A*, **34A**(6), 407-436.

Mayeres, I. and S. Proost (1999). Marginal tax reform, externalities and income distribution. Working Paper, Centre for Economic Studies, Catholic University Leuven. *Journal of Public Economics*, forthcoming.

Mills, D.E. (1981). Ownership arrangements and congestion-prone facilities. *American Economic Review, Papers and Proceedings*, **71**(3), 493-502.

Mohring, H. (1985). Profit maximization, cost minimization, and pricing for congestion-prone facilities. *Logistics and Transportation Review*, **21**, 27-36.

Nijkamp, P. and A.S. Rienstra (1995). Private sector involvement in financing and operating transport infrastructure. *Annals of Regional Science*, **29**, 221-235.

Parry, I.W.H. and A.M. Bento (1999). Revenue recycling and the welfare effects of congestion pricing. Working Paper, Resources for the Future, Washington.

Pigou, A.C. (1920). *Wealth and Welfare*. Macmillan, London.

Richardson, H.W. (1974). A note on the distributional effects of road pricing. *Journal of Transport Economics and Policy*, **8**, 82-85.

Schrank D, and T. Lomax (1999). The 1999 Annual Mobility Report Information For Urban America, Texas Transportation Institute, Texas A&M University System, College Station, Texas http://mobility.tamu.edu.

Small, K.A. (1992a). *Urban Transportation Economics. Fundamentals of Pure and Applied Economics*. Harwood, Chur.

Small, K.A. (1992b). Using the revenues from congestion pricing. *Transportation*, **19**(4), 359-381.

Small, K.A. and J.A. Gómez-Ibáñez (1998). Road pricing for congestion management: the transition from theory to policy. In *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* (K.J. Button and E.T. Verhoef, eds.), pp. 213-246. Edward Elgar: Cheltenham, UK.

Small, K.A. and J.A. Gómez-Ibáñez (1999). Urban transportation. In: *Handbook of Regional and Urban Economics* 3 (P. Cheshire and E.S. Mills, eds.), pp. 1937-1999. North-Holland, Amsterdam.

Small, K.A. and J. Yan (1999). The value of "value pricing" of roads: Second-best pricing and product differentiation. *Journal of Urban Economics*, forthcoming

Smith, T.E., E.A. Eriksson and P.O. Lindberg (1995). Existence of optimal tolls under conditions of stochastic user-equilibria. In: *Road Pricing: Theory, Empirical Assessment and Policy* (B. Johansson and L-G. Mattsson, eds.), pp. 65-87. Kluwer Academic Publishers, Boston.

Tabuchi, T. (1993). Bottleneck congestion and modal split. *Journal of Urban Economics*, **34**, 414-431.

Verhoef, E.T. (1998). Second-best congestion pricing in general static transportation networks with elastic demands. Working paper, Free University of Amsterdam. *Regional Science and Urban Economics*, forthcoming.

Verhoef, E.T., R.H.M. Emmerink, P. Nijkamp and P. Rietveld (1996). Information provision, flat- and fine congestion tolling and the efficiency of road usage. *Regional Science and Urban Economics*, **26**(5), 505-530.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1996). Second-best congestion pricing: the case of an untolled alternative. *Journal of Urban Economics*, **40**(3), 279-302.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1997a). The social feasibility of road pricing: a case study for the Randstad area. *Journal of Transport Economics and Policy* **31**(3), 255-267.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1997b). Tradeable permits: their potential in the regulation of road transport externalities. *Environment and Planning B: Planning and Design*, **24B**, 527-548.

Verhoef, E.T., J. Rouwendal and P. Rietveld (1999). Congestion caused by speed differences. *Journal of Urban Economics*, **45**, 533-556.

Verhoef, E.T. and K.A. Small (1999). Product differentiation on roads: second-best congestion pricing with heterogeneity under public and private ownership. Discussion paper TI 99-066/3, Tinbergen Institute, Amsterdam-Rotterdam.

Vickrey, W.S. (1963). Pricing in urban and suburban transport. *American Economic Review*, **53**, 452-465.

Vickrey, W.S. (1969). Congestion theory and transport investment. *American Economic Review* (*Papers and Proceedings*), **59**, 251-260.

Vickrey, W.S. (1971). Responsive pricing of public utility services. *Bell Journal of Economics and Management Science*, **2**, 337-346.

Viton, P.A. (1995). Private roads. *Journal of Urban Economics*, **37**(3), 260-289.

Walters, A.A. (1961). The theory and measurement of private and social cost of highway congestion. *Econometrica*, **29**(4), 676-697.

Yang, H. (1999). Evaluating the benefits of a combined route guidance and road pricing system in a traffic network with recurrent congestion. *Transportation*, **20**, 299-321.

Yang, H. and H.-J. Huang (1997). Analysis of time-varying pricing of a bottleneck with elastic demand using optimal control theory. *Transportation Research B*, **31B**(6), 425-440.

Yang, H. and H.-J. Huang (1998). Principle of marginal-cost pricing: how does it work in a general road network? *Transportation Research A*, **32A**(1), 45-54.